Published online ahead of print February 1, 2013

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–19
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

informs®

http://dx.doi.org/10.1287/mnsc.1110.1648

Social Ties and User-Generated Content: Evidence from an Online Social Network

Scott K. Shriver
Columbia Business School, Columbia University, New York, New York 10027,
ss4127@columbia.edu

Harikesh S. Nair
Graduate School of Business, Stanford University, Stanford, California 94305,
harikesh.nair@stanford.edu

Reto Hofstetter
Center for Customer Insight, University of St. Gallen, CH-9000 St. Gallen, Switzerland,
reto.hofstetter@unisg.ch

We exploit changes in wind speeds at surfing locations in Switzerland as a source of variation in users' propensity to post content about their surfing activity on an online social network. We exploit this variation to test whether users' online content-generation activity is codetermined with their social ties. Economically significant effects of this type can produce positive feedback that generates local network effects in content generation. When quantitatively significant, the increased content and tie density arising from the network effect induces more visitation and browsing on the site, which fuels growth by generating advertising revenue. We find evidence consistent with such network effects.

Key words: marketing; user-generated content; social networks; monotone treatment response; monotone treatment selection; monotone instrumental variables; homophily; endogenous group formation; correlated unobservables; endogeneity

History: Received June 4, 2010; accepted June 20, 2012, by J. Miguel Villas-Boas, marketing. Published online in *Articles in Advance*.

1. Introduction

We measure empirically the codetermination of user-generated content and social ties in an online social network. Our main point is that online social networks are potentially subject to network effects in *content generation*. Low-level network effects can arise in the following way: increasing a user's social ties on the network may induce him to post more content, which then causes him to obtain more ties, which in turn causes him to post more content and so on. This process generates local positive feedback between ties and content, which thereby produces a self-reinforcing "virtuous cycle." Such feedback may be linked to several motivations. Altruistic user intentions (users wanting to share information with others) and a desire to increase social status within a peer group may cause users to post more content when they have more ties (an audience effect). Users who post more content may in turn receive more ties because of their popularity or the intrinsic value others derive from consuming their content. Positive feedback of this sort may be an important component of an online network's revenue model. When quantitatively significant, the increased content and tie density arising from the network effect induces more visitation and browsing on the site, which fuels growth by generating advertising revenue. The academic literature on social networking (reviewed later) has paid relatively less attention to such content-related network effects. The main goal of this paper is to investigate such network effects using data from a real-world social network.

Empirically analyzing these network effects is complicated by identification problems endemic to the measurement of social effects from observational data (e.g., see Manski 2000). Formally, the network effect can be represented by a simultaneous equations model in which content drives social tie formation and tie formation in turn drives content generation. Positive feedback exists when content has a positive effect on the propensity to obtain more social ties *and* when having more ties has a positive effect on the propensity to post content. Unfortunately, identifying causality in these two effects is confounded by the simultaneity of their coproduction and by the fact that, in addition to the causal effects of interest, several other sources of spurious correlation can explain the covariation in content and ties observed in social network data.

This paper presents data from an online social network that documents the codetermination of content and ties and discusses some strategies to address this identification problem. The success of these strategies is dependent on primitive assumptions about behavior that we discuss in the paper. Some of the identifying assumptions may be violated under alternative stories of user behavior, and therefore our results should be seen with this caveat in mind. We first consider the problem of measuring the effect of content generation on the propensity to obtain social ties. To address the simultaneity of coproduction, we look for a source of variation in users' propensity to post content on the website that may not directly affect tie formation. To do this, we exploit institutional features of our empirical setting. Our setting involves the content posting and tie formation activities of users of an online social network named Soulrider.com, which hosts one of the largest sports-based online communities in Switzerland. We focus on the large community of windsurfers on the site. Users post content about their surfing activities, including blogs and reports of wind speeds and conditions at specific surfing locations. Our data comprise historical information on content posted and social ties formed by these users on Soulrider.com. We augment the website data with high-frequency information on wind forecasts at all surf locations in the country obtained from the Swiss Meteorological Office (http://www.meteoswiss.ch). We utilize the fact that surfing is possible for most users only if wind speeds are greater than or equal to 3 BFT.[1] We show that wind speeds significantly explain the observed variation in content postings on the website. We use the variation in wind speeds at various lake locations frequented by users as shifters of users' propensity to visit surf locations and to subsequently post content about their surfing activity. Additionally, we exploit the panel nature of the data to include user and time-period fixed effects to control for spurious sources of correlation that drive content generation and tie formation.

We then discuss the measurement of the reverse effect—the effect of social ties on the propensity to post content. The main challenge here is that those that tend to obtain more social ties for unobservable reasons (e.g., "gregariousness") may also post more content for the same reasons. Hence, the number of tie requests received is econometrically endogenous. This is a manifestation of the "homophily" or "endogenous group formation" problem in measuring social interaction (e.g., Moffitt 2001). We discuss a strategy using the number of friend requests to an agent's friends as an instrument for his friend requests, including individual and time fixed effects, to control for agent-specific and common unobservables that may drive friendship formation. Intuitively, we exploit variation in network position under the assumption that after controlling for any common shocks, an agent's content generation is likely driven only by the environment facing him, and not his friends. These instrumental variable (IV) assumptions can be violated if wind affects friendship formation directly because users meet future online friends at more windy surf locations, or if friends receive tie-formation requests in a way that is directly related to the focal agent's blogging activity. If these exclusion restrictions are violated, our estimates should be interpreted only as correlational and not causal. We provide some evidence suggesting that the exogeneity assumption is reasonable but do not rule out exogeneity concerns fully. We also present estimates from two alternative approaches that use much weaker assumptions and deliver bounds on the treatment effects.

For the bounds, we use an incomplete model of the social network that leaves the formal model of tie formation and content generation unspecified but incorporates the fact that ties and content are not randomly allocated across units of observations. Our approach is in the spirit of Haile and Tamer (2003) and Ciliberto and Tamer (2009), in the sense in that we recognize tie formation is nonrandom and jointly determined with content generation, but we leave the actual model-generating ties unspecified. Following the framework of Manski (1997), we impose weak monotonicity assumptions on the response of content to ties and of ties to content that imply informative, nonparametric bounds on the causal effects of interest. Following Manski and Pepper (2000), we also report bounds under the assumption that wind speeds and friends' friend requests are monotone instrumental variables (MIVs), a weaker assumption than the requirement that they are IVs for content and friends, respectively. Both bounds are robust to selection and to misspecification concerns. We find the point estimates from the IVs are contained in the bounds we obtain and show that monotonicity is not rejected by the data.

An alternative solution to the codependence of network formation and content generation is to specify a formal model for tie formation and content production and to incorporate into estimation the likelihood of these actions for all network members. However, one difficulty of this approach is that the endogenous actions of tie formation and content posting are the equilibrium outcomes of a complicated, multiagent network game, which needs to be solved for every guess of the underlying parameter vector to compute the likelihood for inference. Apart from the computational complexity, this estimation is also subject to

[1] BFT stands for "Beaufort," the international wind scale used in weather reporting.

misspecification bias if the wrong network model is assumed. More difficulties arise from the fact that realistic models of network formation games that can explain features of observed online network structure also tend to be prone to multiple equilibria (e.g., see Jackson 2008). The bounds strategy avoids taking a stance on a specific network game but recognizes the codependence in inference and thus makes some headway on this difficult inference problem.

Our analysis provides evidence consistent with network effects. We find that a user's social ties are positively associated with his content generation and that content generation in turn is positively associated with obtaining social ties. We find large effects of controlling for endogenous group formation and correlated unobservables. In the absence of these controls, we find that measures of social effects are overstated about 25%–30% on average across model specifications. To interpret our estimates, we calibrate the revenue implications of content generation to the network. Translating content to advertising revenues, we calculate an incremental blog is worth about 0.736 Swiss francs (CHF) on average to Soulrider.com and that local feedback with ties can augment the revenue by about an additional 5.5% on average across users.

Our paper relates to extant literature investigating the process of tie formation in online social networks and the influence of user-generated content on online activity.[2] Recent studies in marketing include Trusov et al. (2009) on online opinion leader identification, Stephen and Toubia (2010) on the economic effects of social commerce, and Mayzlin and Yoganarasimhan (2012) and Katona and Sarvary (2009) on link formation. A growing stream of research also explores the effect of user-generated content, broadly defined, on economic outcomes. These include Dhar and Chang (2009) on the effect of blog posts on future music online sales, Dellarocas (2006) on Internet forums, Duan et al. (2008) and Chintagunta et al. (2010) on the effect of online reviews on movie box office performance, Chevalier and Mayzlin (2006) and Oestreicher-Singer and Sundararajan (2012) on the effect of book reviews on online book sales, Albuquerque et al. (2012) on the interdependence between creating and purchasing online content, and Zhang and Sarvary (2011) on competition with user-generated content. These studies suggest that user-generated content plays an important role in consumer decisions and competition but is not concerned with its interaction with social ties per se. A parallel stream of literature describes motives of users for generating content. Ahn et al. (2011) and Kumar (2011) develop structural models of content creation and consumption, and Zhang and Zhu (2011) demonstrate that content creation responds to audience size. Nov (2007) surveyed contributors to Wikipedia.com and found that fun and ideology are the primary drivers of content generation. Hennig-Thurau et al. (2004) report that consumers' desire for social interaction, desire for economic incentives, concern for other consumers, and the potential to enhance their own self-worth are the primary factors motivating consumers to articulate themselves online. Their finding that desire for social interaction is a driver to post content has been corroborated by descriptive studies like Bughin (2007), which documents the desire for fame and to share experiences with friends as reasons for posting content online. Likewise, Nardi et al. (2004) emphasize the community aspect of content creation. A survey we conducted finds motivations that are consistent with these findings. Although this stream of literature suggests the existence of effects between content and social ties, it does not investigate the codependence explicitly.

To date, few studies have explicitly discussed or investigated the potential interplay between user-generated content and social ties. Godes et al. (2005) include in their definition of social interactions the possibility of non-face-to-face interactions, such as passive observations that could include effects of endogenously generated content on social ties. Ghose and Han (2011) state that little is known about how content creation is related to content usage or social network characteristics. Goldenberg et al. (2009), Katona et al. (2011), and Yoganarasimhan (2012) document that network structure matters in the diffusion of content. Narayan and Yang (2007) model the tendency to form ties, but do not consider the role of content. Using data from Wallop, an online personal publishing and social networking system, Lento et al. (2006) test how the number and nature of social ties are related to people's willingness to contribute content to a blog. Their aim is to predict future content contribution, not to measure causality; hence, the study does not control for the endogeneity concerns we outline here. Their results show that the particular type and strength of the tie are correlated with the effect of social ties on content. In their descriptive analysis, however, Lento et al. (2006) find that the number of in-degree ties is not a significant predictor of future content production. Hence, we believe this is one of the first studies to measure the codetermination of content and ties while considering the identification challenges implied by simultaneity, endogenous group formation, and spurious correlation caused by common unobservables.

[2] More generally, this paper relates to the large literature on social interactions (see Hartmann et al. 2008 for a recent survey).

The remainder of this paper is organized as follows. In §2, we provide a short overview of the online social network Soulrider.com, from which the data are collected. In §3, we present a variety of estimators for the effects. In §4, we present the results and the implications of our model. In §5, we conclude the paper.

2. Empirical Setting

We now provide a brief overview of the empirical setting for our data: an online social network named Soulrider.com. We discuss how we operationalize the key variables of interest in the data. Subsequently, we discuss the model and empirical strategy for identification of the causal effects of interest. A more detailed description of the data set is provided later with the results.

2.1. Soulrider.com

Soulrider.com is a privately held community website based in Europe that focuses on extreme sports such as windsurfing, surfing, and snowboarding. The site was launched in 2002 and as of June 2011 had a total of 10,677 registered users.

Users of Soulrider.com can both consume existing content and submit their own. Most content on the website, such as blogs or forum messages, is generated by users themselves. Other content, such as national-level sport industry news, is provided by third-party contributors and is not affected by users. Users who wish to post content or engage in social networking activities are required to create a free account on the website. The main value of content is *informational*. Blogs and posts typically contain information about where surf conditions are best, where other users plan to surf in a given week, and reports on wind speeds at those locations. These posts help users better organize and coordinate their sporting activities.

In addition to consuming and generating content, users can create ties to other users and thereby take part in an online social network. Tie formation among users is facilitated by the website through various functions such as a people search engine, an email invitation tool, interest groups, and an "add as friend" function that handles the mechanics of creating ties in the online interface. Communication among users is eased by internal mail functionality, instant messaging, and public chat. The process of "friending" (or tie formation) on the site involves sending a tie-formation request through the website that has to be accepted by the recipient. Friending enhances access to content. By default, users on Soulrider.com can see others' profiles and content. However, access to one's profile and albums and the ability to contact via email may be restricted by some users to only their friends. Friending also provides benefits to users in obtaining status updates and enhanced online interaction with others. Upon login, friends who are currently online are displayed on the welcome screen for a user, and direct access is provided to all their online sessions and activities. Users can also add comments and tag photos of only those who have accepted them as friends. This is the norm in most online social networks.

Soulrider.com is representative of networks targeting young adults. As of June 2011, 75% of the users on the website were male and 25% were female. The mean user was 29 years old and possessed 1.4 friends. The social network grew by 1,164 (12.3%) users from June 2010 to June 2011. Within this period, 470 (4.4%) added at least one friend, which generated 1,193 new social ties; blogs were contributed by 477 (4.5%) users. The site averaged 37,211 unique visits per month, with visitors staying on the website for an average of 3.8 minutes, generating a monthly average of 330,108 page impressions. Google has indexed 31,800 pages and 27 backlinks for Soulrider.com, leading to a pagerank value of 6.[3]

Although it is smaller than some well-known online social networks, content generation on the site is similar in pattern to those reported at larger online social networks. Content contribution on Soulrider.com follows the so-called participation inequality or "90-9-1" principle (e.g., Brothers et al. 1992, Ochoa and Duval 2008), a commonly observed empirical pattern of online user-generated content distributions, whereby 90% of users form the "lurkers," or audience, who do not actively contribute to the site, 9% of users are "editors," sometimes modifying content or adding to an existing thread but rarely creating content from scratch, and 1% of users are "creators," driving large amounts of the social group's activity. The 90-9-1 principle has been used as a rule of thumb to characterize user-generated content production and roughly holds for Soulrider.com as well, strengthening to some extent the external validity of our findings.[4] To summarize, Soulrider.com is a medium-sized social network that appeals to a core community with shared interests, grows primarily by word of mouth, and generates content in a fashion representative of many online social networks.

2.2. Data and Variable Operationalization

We worked with Soulrider.com to add a logging functionality to capture details of network growth and

[3] Based on impression statistics from Google Analytics for the 30-day period ending June 16, 2011.

[4] For instance, of 22,279 unique visitors to the site in December 2010, 2,161 (9.7%) produced content. Also, 1% of registered users are heavy users, accounting for 47% of all produced content (blogs).

content generation and consumption. Our data comprise complete details of user tie formation and content generation from March 2, 2009, to November 7, 2010, a period spanning 89 weeks. For the purposes of this paper, we focus on a group of 703 self-identified windsurfers on the website, giving us a panel with 57,040 user-week observations for analysis. The large panel size facilitates the inclusion of fixed effects to control for unobservables, as well as relaxing several parametric assumptions for inference.

We operationalize user content via a generic variable we call *blogs*, which counts the number of postings a user has made to the website each period. A posting is counted as adding 1 to the *blogs* variable if it contains any text or photos. Thus, if a user posted a photo and a text message on the website, two text messages, or two photos, the *blogs* variable would equal 2. One limitation of this approach is that we focus only on the *incidence* of postings. We do not account for the type of content added because of the difficulties and ambiguity associated with sorting posts into predefined classes. We operationalize ties by a variable, *friends*, which counts the new friendship ties requested by others to each user per period. Essentially, we ask if content creation facilitates new friend requests and whether new friend requests in turn facilitate the creation of new content.[5]

2.2.1. Understanding Motivations and Mechanisms: A Survey. To learn about the mechanisms behind blogging and link formation, we implemented an extensive survey on Soulrider.com. We used an open-ended survey to allow users to freely express their motivations for friending and posting content. We then used content analysis involving pattern matching to analyze the qualitative response data. We categorized the motives as stated by users into predefined types and counted frequencies of occurrences of these categories. This exploratory analysis helped us set up the model, understand user behavior and mechanisms, and interpret the data.[6]

We invited 3,157 self-identified windsurfers of Soulrider.com by email to participate in the survey. Two hundred sixty-six (8.43%) chose to participate. In the survey, we asked the following three open-ended questions related to blogging: Q1: “What are and have been reasons and motives for you to add a windsurf-session on Soulrider.com?” (see Table A1 in the online appendix). Q2: “If you are online on Soulrider.com, what specifically prompts you to insert a windsurf-session?” (see Table A2 in the online appendix). Q3: “What benefit does Soulrider.com provide relating to your windsurf sessions?” (see Table A3 in the online appendix). Similarly, we asked three open-ended questions related to friending: Q4: “What are and have been reasons and motives for you to link to somebody as your friend on Soulrider.com?” (see Table A4 in the online appendix). Q5: “If you are online on Soulrider.com, what specifically prompts you to send a friendship request to somebody?” (see Table A5 in the online appendix). Q6: “What benefit do you obtain from your friends on Soulrider.com?” (see Table A6 in the online appendix). For each question, respondents were advised to answer with short text or a few key words. For further analysis, we only used answers to questions Q1–Q3 from 121 respondents who indicated having at least one friend on Soulrider.com and answers to questions Q4–Q6 from 89 respondents who reported posting at least one blog.

We then processed the survey data using a content analysis protocol (e.g., Krippendorff 2004). Each statement was coded using a coding schema. Following suggestions in the literature on textual coding (Corbin and Strauss 2008), we started with an open coding of the first 50 statements to derive descriptive codes that were close to the actual text and then grouped codes into categories. In second-stage coding, we applied this scheme to all the data to tabulate frequencies of occurrences per code and category. All coding was performed using the software package ATLAS.ti (Friese 2011). The data analysis was done by several coders and checked for intercoder reliability. Reliability of these codes was then assessed using two independent raters, who recoded the data into the given code categories. These raters were not current users of Soulrider.com and were not familiar with the details of the research project. Interrater reliability values (Cohen’s Kappa) of 0.65 on average indicated good reliability of the derived categories.

Determinants of Blogging. Fifty-five percent of users in the survey stated “altruistic” reasons for blogging on Soulrider.com (see Table A1). Altruistic users express a need to inform others about their experiences and to share photos and various kinds of information related to surfing or conditions. In contrast, 18% of users indicated they blog only for their own purpose (egocentric motivation). This includes keeping a diary or archive of one’s surf sessions, one functionality offered by the website. A substantial number of users indicated that they use their blogging to express or promote themselves (20%), which is not

[5] We do not answer the separate question of how content generation is affected by the number of friend requests a user *initiates*. This is a different economic problem in which each user chooses blogging frequency and outgoing friend requests jointly to maximize his utility, which requires specification of a formal structural model of the network game. This problem is outside the scope of the current analysis and is the subject of future research.

[6] The coding scheme, sample quotes, definitions, reliability measures, and frequencies are available by request from the authors and in the companion online appendix (http://faculty-gsb.stanford.edu/nair/documents/TechnicalAppendix.pdf).

surprising, because many professional windsurfers use the site to appeal to potential sponsors. To investigate the determinants at a more detailed level, we asked users what prompts them to post a blog (see Table A2). A majority (62%) indicated that they would add a new entry shortly after they went surfing. Others (22%) stated they would not blog about every session, but only about those that were special or where conditions were particularly unique (e.g., wind at storm level). Interestingly, 16% also explicitly indicated that they would blog to facilitate interaction or socializing with others. As meeting or communicating with other surfers at the spot is often perceived to be difficult, some use blogging as a means to find out which other surfers surfed at the same spot on the same day. Socializing is also a major value driver of blogging for some users (22%) (see Table A3, Q3). Information motives, either at the community (altruistic) or individual (egocentric) level, are the most frequently stated drivers of the value of blogging (these categories together account for 63% of the statements).

Determinants of Friending. The stated reasons for friending are more fragmented (see Table A4). We grouped descriptions into four main categories. Thirty-four percent of users indicated that they would link to somebody only if he or she was also a close friend in real life (i.e., a strong tie). Others stated they would also link weaker contacts, as long as they both shared common interests (21%). Twenty-seven percent of users state they intend to keep in touch with their friends via the website. Consuming and sharing content, such as blogs, photos, or information about surf spots or surf destinations, also drove friending (18%). Answers to Q5 provide more detailed insights into the specific determinants of sending friend requests. As before, users are likely to send a friend request if they observe the online presence of one of their close friends on Soulrider.com (39%). Other determinants include "keeping in touch" (16%), consuming and sharing content (6%), and networking with like-minded people (38%). Some users indicated that they friended somebody they recently met offline (e.g., at surf sessions, on holidays, or when trading windsurf equipment), though this proportion is small (9%). Overall, the value of friending seems to be mostly related to consuming and sharing friend- or sport-related content (61%). Twenty percent of users state that the website facilitates communication between users. Interestingly, a small proportion (11%) specifically indicated that they derive no value from their friends.

We find wind speed is strongly associated with blogging. In answering Q2, 62% of the respondents indicated that they would post a blog after they went surfing or have new content related to a surf session. In addition, 22% indicated they would blog if they experienced a unique session. Finally, we use these data to assess informally whether surfing at the same spot on the same day is a driver of friending. Counting the incidence of statements like "to link to somebody I know from surfing" or "to link to somebody I met at a surf session" as reasons for friending, we find these account for 11 quotations or 7.14% of all quotations in this coding scheme altogether (see Table A5). Although small, this response rate shows that we cannot rule this motivation for link formation out completely, and thus our identification strategy should be seen with this caveat.

3. Empirical Approach

We now discuss our empirical strategy. The goal of the empirical work is to learn the effects of *blogs* on *friends* and of *friends* on *blogs*. In §3.1, we discuss how we identify the effect of *blogs* on *friends*. In §3.2, we discuss identification of the reverse effect.

The plan of this paper is to first establish evidence of a correlation between *blogs* and *friends* in the data. We then show that this correlation is robust to the inclusion of individual fixed effects (to control for homophily) and of time period fixed effects (to control for correlated common unobservables). Interpreting these correlations as causal effects requires additional assumptions. We present an IV approach based on wind speeds, which delivers estimates of causal effects under the assumption of exogeneity of these instruments. If these assumptions are violated, the causal interpretation no longer holds. We present some evidence suggesting that the exogeneity assumptions are reasonable, though we cannot conclusively establish their validity a priori. We also present bounds on causal effects that rely on weak monotonicity assumptions. We find that the point estimates from the IVs are contained in the bounds we obtain.

3.1. Effect of Content Generation on Tie Formation

We start with a linear model linking *blogs* (b_{it}) and *friends* (f_{it}). Here, i stands for individual and t stands for week. Let $\mathcal{B}_{it}$ and $\mathcal{F}_{it}$ denote the cumulative number of blogs written and friendship requests, respectively, to agent i at the beginning of (but excluding) period t. The linear model is

$$f_{it} = \alpha_{1i} + \gamma_{1t} + \theta_1 b_{it} + \lambda_1 \mathcal{B}_{it} + \zeta_1 \mathcal{F}_{it} + \delta_1 z_{it} + \varepsilon_{1it}. \quad (1)$$

Equation (1) is in essence a dynamic panel data model in which we allow past blogging ($\mathcal{B}_{it} = \sum_{\tau=-\infty}^{t-1} b_{i\tau}$) and past friendship requests ($\mathcal{F}_{it} = \sum_{\tau=-\infty}^{t-1} f_{i\tau}$) as well as current blogging to affect current friendship requests. Equation (1) also allows *friends* to depend on other exogenous variables (z_{it}); in our empirical specifications, we include in z_{it} the number of days a user has been registered on the site as a control for common tieing trends over user tenure. The main effect of interest for our study is θ_1, the marginal effect of *blogs* on *friends*.

A concern with estimating θ_1 is that individuals are sorted into groups on the basis of unobserved tastes and that these tastes may also drive blogging behavior. For instance, we may suspect that unobserved factors like "popularity" that cause a specific agent to receive more friendship requests also cause him to post more content. This is a version of the endogenous group formation problem arising from "homophily," the tendency of agents with similar tastes to form social groups. Homophily is a pervasive feature of social networks (e.g., see McPherson et al. 2001) and has been shown to be empirically important in online social network data (Ansari et al. 2011, Braun and Bonfrer 2011). One solution to the endogeneity induced by homophily is facilitated by the availability of panel data. Assuming agent tastes are time invariant, one can control for time-invariant aspects that drive endogenous group formation via agent fixed effects, α_{1i} (e.g., Narayanan and Nair 2013).

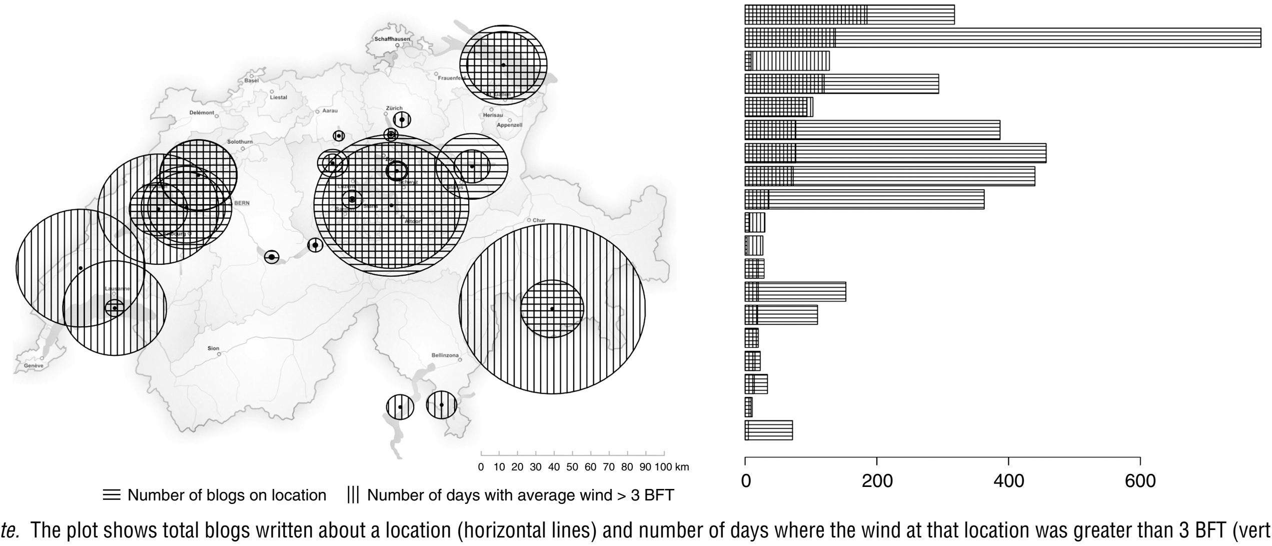
In addition, one may be concerned that comovement in *friends* and *blogs* across users and time may be generated by common, time-varying unobservables. For instance, agents may be significantly motivated to request ties and post content related to surfing during periods particularly conducive to surfing activity (e.g., great weather) or when surfing is more salient than usual in sports enthusiasts' minds (e.g., a large surf tournament occurred). These common shocks, when unobserved by the analyst, are a source of spurious correlation. The time period fixed effects, γ_{1t}, control for these kinds of concerns.

Even after employing these controls, concerns may continue to persist about the identification of θ_1. First, there may be time-varying individual-specific unobservables driving friendship requests that may be correlated with blogging propensity. For instance, an unobserved (to the econometrician) local event featuring agent i's recent windsurfing activity may cause a specific subset of users to request friendship ties with that user, and that event may also cause agent i to blog more on Soulrider.com. In this story, the source of spurious correlation is individual and time specific, which is not picked up by fixed effects that are common across agents for a given week or common across weeks for a given agent. Second, b_{it} may directly be a function of f_{it} in an equation analogous to (1) that determines blogging activity. This situation could arise, for instance, if the true data-generating process is one in which some individuals post content primarily as a way to obtain more friends on Soulrider.com. To address these concerns, we look for a source of variation in *blogs* (b_{it}) that can be excluded from the equation determining f_{it} and thereby serve as an instrument for b_{it}.

Instruments for Blogs: *Variation in Wind.* Our strategy exploits differences in wind speeds across surf locations in Switzerland. We expect content to be produced concomitant with actual surfing activity, which is in turn driven by wind speeds at surf locations. If wind speeds affect blogging, we would expect to see that blogs relating to a particular location are more likely when wind speeds there are higher. We present geographic plots and statistical tests showing this is true in our data.

Figure 1 plots the geographic distribution of blogging activity and wind. We plot with horizontal lines the number of total blogs written about a particular

Figure 1 Evidence of Positive Covariation of Blogging and Wind Speed



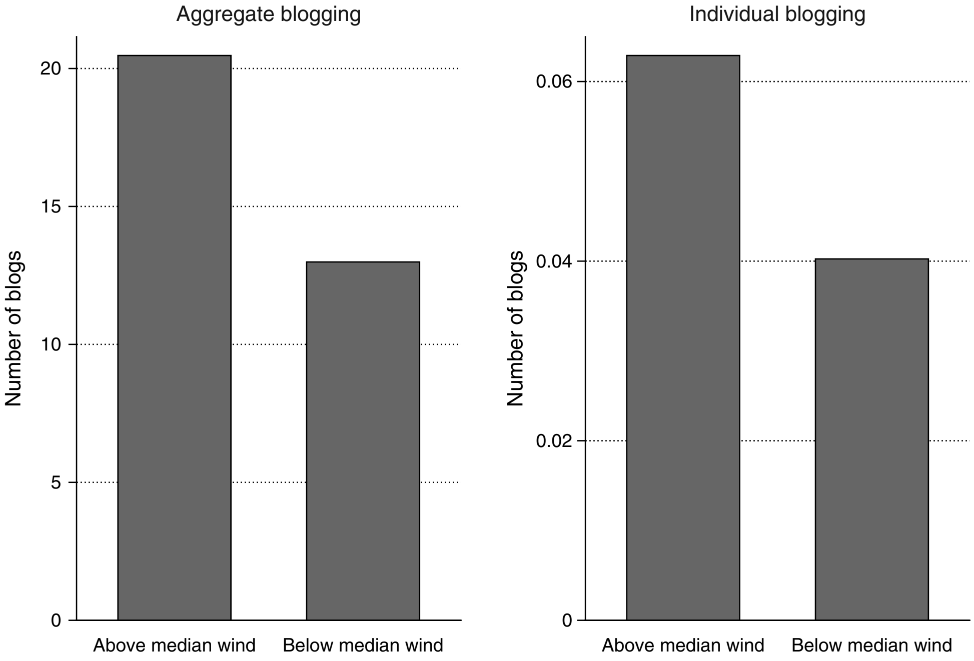
Note. The plot shows total blogs written about a location (horizontal lines) and number of days where the wind at that location was greater than 3 BFT (vertical lines).

location (lake) and with vertical lines the number of days where the wind at that location was greater than 3 BFT (left). The barplot (right) displays the same information. Each bar represents 1 of the 19 surfing locations. The bars are sorted by the number of surfable days in descending order. The number of blogs and surfable days at a specific location can be read from the x-axis. We see there is evidence of a strong correlation, implying that blogging responds to wind speeds at the focal location (corr = 0.39, $t = 3.90$, $p < 0.001$). As a test, we compute the mean weekly wind speed at every location and create a low- and high-wind-speed group by splitting at the median. We test whether the mean (across weeks) of total weekly blogging is statistically significantly different between the low- and high-wind-speed groups. The mean number of weekly aggregate blogs in the low-mean-wind group is 12.99 and in the high-mean-wind group is 20.47. Means differ significantly between the two groups ($t = 3.23$, $p < 0.002$).

To also see whether wind drives blogging at the individual-week level, on the right-hand side of Figure 2, we plot a median split of wind speed at the preferred location at the individual-weekly level and compare the per-user weekly blogging intensity between the two groups. We see individual blogging is significantly higher in the high-wind-speed group ($t = 9.25$, $p < 0.001$). On the left-hand side of Figure 2, we also plot the mean (across weeks) of the aggregate (across users) weekly blogging between the same two groups. We see aggregate blogging about the preferred location is also significantly higher in the high-wind-speed group ($t = 3.23$, $p < 0.001$). Finally, to check whether users who generate content about a particular location are also those who have a large number of friends, in Figure 3 we plot blogging arising from a user overlaid with his social connections. On the right of the figure, each line in the barplot represents one particular user of the social network. Users are sorted (in descending order) by their number of social ties (black lines). The grey lines indicate the number of blogs per user. The figure provides evidence for a robust correlation (corr = 0.41, $t = 12.01$, $p < 0.001$), as more blogs about a location are generated by users who have more ties.

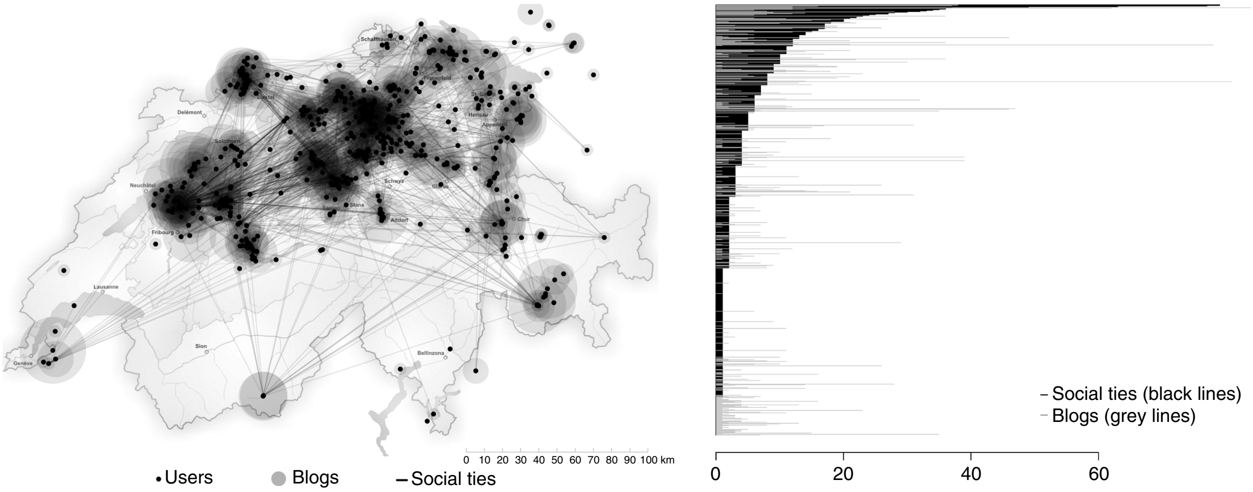
These plots evince patterns that suggest blogging is linked to wind speeds and that individuals who blog often also tend to be well connected. Our identification strategy is tied to using wind speeds as exogenous shifters of blogging, which are excluded from the propensity to have friends. The assumption behind this argument is that users do not form friendships at surfing locations per se, so a higher wind speed does not *directly* cause a user to have more friends. We believe this assumption is not unreasonable based on our results from the survey and the nature of the sport. Friending on the website is primarily driven by prior offline ties and from seeing

Figure 2 Blogging Incidence in Relation to Wind Speed



Notes. Left panel: Mean total weekly blogging split (across all user-weeks) at the median wind speed at preferred surf location. Right panel: Mean weekly blogging per-user split (across all user-weeks) at the median wind speed at preferred surf location.

Figure 3 **Evidence of Positive Covariance of Blogging from a User with His or Her Network Ties**



Notes. Left: Geographic plot centered on each user. Right: Black lines represent users, sorted by number of ties; grey lines denote blogs per user.

profiles on the site, not by meeting on the beach, which is typically a short interaction (our survey suggests ties result from multiple/repeated interactions and not one-time meetings). For instance, in Figure 3 we do not see much evidence for geographic proximity determining the distribution of ties, as would be the case if users were primarily connecting with each other following their visits to nearby surf locations.[7]

As a robustness check of the assumption, we investigate whether average friendship requests per user are higher during weekends—when users may be more likely to socialize at the beach—than on weekdays.[8] We find little difference (0.004 during weekdays versus 0.003 on weekends). The raw correlation between blogs and friends at the day-user level is 0.04 during the weeks and 0.014 during weekends (going in the other direction of the socialization story—higher correlation during weekdays).

As a stronger test, we look for alternative model-free evidence that friendship formation does not occur in direct response to changes in wind conditions. If high wind makes surfers more likely to go to a lake and more likely to meet future online friends there, we would expect to see more links formed between a user and those who live in closer geographic proximity to his preferred surf location. Consider an agent i who receives friendship requests and the set of agents j $(1, \ldots, \text{number of agents} - 1)$ who send the requests. Let $D(i, j)$ be the distance from j's home location to i's preferred surf location. If wind leads to friendship, the propensity of i to form a link with j should be a decreasing function of $D(i, j)$—if there is no such geographic dependency, it is unlikely that friendship formation is directly influenced by wind conditions.

For each user pair (i, j) in our data, we compute $D(i, j)$ as the driving distance between agent j's home and i's preferred surf location in kilometers using the Google Maps application programming interface (we compute 490,976 such pairs). We then run logit models of the effect of $D(i, j)$ on the propensity of all pair of users i and j to become friends on Soulrider.com. Following the survey evidence, we include proxy variables for the chance that users know each other from before in real life, as well as a variable measuring the salience of the online presence of those they know on Soulrider.com. To capture the first factor, we compute the distance between agents' physical homes—the assumption being that two agents who live close to each other are more likely to know each other offline (people who live close to one another are more likely to meet, regardless of wind conditions). To capture the second factor, we use the Web logs from the site and create a variable measuring the duration in seconds when both agents i and j were online at the same time on Soulrider.com prior to their friendship on the site. If users are online at the same time, they are more likely to observe each other because the site displays a visible list of all users currently logged in (this is similar to other social networks like Facebook). Table 1 reports these regressions.

Consistent with the survey, we find that users who live closer to each other and those who were frequently online on Soulrider.com together are more

[7] High winds at a surfing location may make a user more interested in friending locals to learn about the surfing conditions around them (e.g., through email). We found that only 16 of 703 (2.3%) users restricted Soulrider.com's email functionality to their friends. Hence, the majority of users could be reached by Soulrider.com email without the need to link them as friends. As a result, anecdotally, we do not believe that the need to communicate is a major driver of friending.

[8] We thank an anonymous referee for this suggestion.

Table 1 Testing for the Mechanisms of Link Formation

Variable	Parameter	t-stat.	Parameter	t-stat.
D(*i*, *j*)	4.68E−04	0.99	6.57E−04	1.38
Distance between homes of agents i and j	−8.94E−03	−16.78	−9.22E−03	−17.05
Duration of overlapping website visits (sec)			6.94E−06	6.15
Constant	−4.67	−113.79	−4.69	−112.37
Observations	490,976			

Notes. The dependent variable equals 1 if users *i* and *j* are friends, 0 otherwise, estimated across all (*i*, *j*) pairs in data. Two binary logit model results reported (with increasingly stringent controls). If high wind makes surfer *i* more likely to go to a lake and to meet future online friend *j* there, we expect Pr(*i*, *j* form a link) is decreasing in *D*(*i*, *j*), the distance between *i*'s favorite surf location and *j*'s home location. Results above show *D*(*i*, *j*) is not a significant predictor of link formation, suggesting this is not the case. Results do not change with a linear probability model, including fixed effects for *i*, for *j*, or for both *i* and *j* (to control for homophily).

likely to be friends, but we do not see any statistically significant effect that users are more likely to form friendships with those that live closer to their favorite surf location ($t\text{-stat}_{D(i,j)} = 1.38$). These effects do not change in a linear probability model that includes fixed effects for i or for j or for both i and j (to control for homophily). This suggests that meeting at the beach is not first order in driving the link formation. Finally, to inspect these relationships visually, in Figure 4 we plot the probability of users i and j forming a link on the y-axis against $D(i, j)$ on the x-axis, conditional on the distance between users' homes. Also included are the best fit lines. Consistent with the regression results, we do not see the negative relationship that we would have expected if people indeed form friendships with those who live closer to their favorite surf location. Given these findings, we think using wind as an IV is not unreasonable.

To operationalize the instrument, for each user i, we compute a variable, $wind\_mean_{it}$, as the weekly mean of the wind speed at his most preferred surfing location.[9] We also compute as $wind\_sd_{it}$ the standard deviation of the weekly wind speed at the same location. We hypothesize that blogging activity will be greater, ceteris paribus, in weeks with high average wind speed and high wind variability, because both the salience and extent of relevant information about surfing conditions will be heightened during such periods. Table 2 reports descriptive statistics of these two variables. We collect both in an instrument vector, $m_{it} = \{wind\_mean_{it}, wind\_sd_{it}\}$.

3.1.1. Estimation. We can now estimate Equation (1) by the generalized method of moments (GMM) using m_{it} as instruments for blogging. In estimation, we need to handle a concern that, including individual fixed effects (α_{1i}), leads to an econometric endogeneity problem when the "stock" variables $\mathcal{B}_{it}$ (cumulative blogs) and $\mathcal{F}_{it}$ (cumulative friend requests) are included, because the presence of stock variables violates the "strict exogeneity" assumption required for the estimation of panel data models with fixed effects—namely, that $E[\epsilon_{1it}\mathcal{F}_{i\tau}] = 0$ and $E[\epsilon_{1it}\mathcal{B}_{i\tau}] = 0$ for all t and τ (see, e.g., Wooldridge 2002). We follow Chamberlain (1992) and estimate the model under the standard, weaker assumption of sequential exogeneity: that $\epsilon_{1i\tau}$ is uncorrelated with current and prior values of the stock variables ($E[\epsilon_{1it}\mathcal{F}_{i\tau}] = E[\epsilon_{1it}\mathcal{B}_{i\tau}] = 0$, $\forall\, \tau \leq t$), which is a natural fit with our model specification. Under the sequential exogeneity assumption, two period or greater lags of $\mathcal{F}_{it}$ are available as instruments for $\Delta\mathcal{F}_{it}$ (as are any function of these variables), with similar relations holding for $\mathcal{B}_{it}$ (Wooldridge 2002). Because a large number of overidentifying restrictions can result in an estimator with poor finite sample properties, we limit the set of instruments for $\Delta\mathcal{B}_{it}$ and $\Delta\mathcal{F}_{it}$ to the vector $q_{it} = \{\mathcal{B}_{i,t-2}, \mathcal{F}_{i,t-2}\}$. After transforming by first differencing, the moments used for estimation become $\mathcal{M}_1 = Q(\Delta f_{it} - (\theta_1 \Delta b_{it} + \lambda_1 \Delta\mathcal{B}_{it} + \zeta_1 \Delta\mathcal{F}_{it} + \delta_1 \Delta z_{it}))$, where Q is the vector $[m'_{it}\, q'_{it}\, z'_{it}]'$, and indicator variables capturing the γ_{1t} parameters have been absorbed into z_{it}. The estimated parameters minimize the GMM objective function, $\mathcal{M}'_1 \mathbb{W} \mathcal{M}_1$, where $\mathbb{W}$ is a weighting matrix. Following Hansen (1982), the optimal $\mathbb{W}$ is inversely proportional to the variance of the moments. We estimate $\mathbb{W}$ via the usual two-step procedure, assuming independent moments in the first step and using the first-step estimate to construct the sample analog of the optimal weighting matrix in the second step.

Table 2 Descriptive Statistics of Model Variables

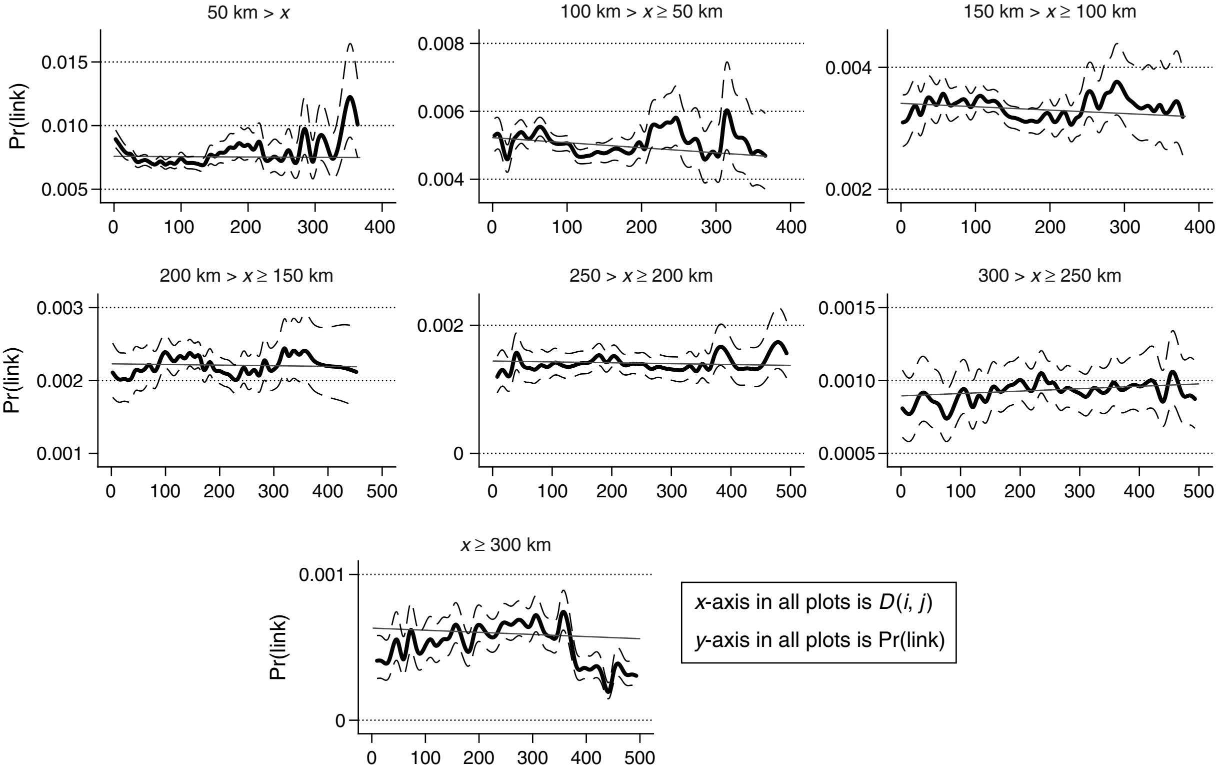
Variable	Obs.	Mean	Std. dev.	Min	Max
blogs (*b*)	57,040	0.051	0.293	0	7
friends (*f*)	57,040	0.026	0.179	0	6
prior_blogs ($\mathcal{B}$)	57,040	4.814	6.536	0	61
prior_friends ($\mathcal{F}$)	57,040	9.716	21.007	0	192
days_online	57,040	909.137	567.310	0	2,535
wind_mean	57,040	2.252	0.444	1.500	4.250
wind_sd	57,040	0.594	0.283	0	4.010
fof_mean	57,040	0.430	1.093	0	18

3.2. Effect of Tie Requests on Content Generation

We now discuss identification of the reverse equation, the effect of *friends* on *blogs*. Estimation of this effect is subject to similar considerations as in the previous case, requiring access to some exogenous variation in *friends* that can be excluded from the *blogs*

[9] Users typically frequent their most preferred surfing locations. Our data reveal that 98.7% of all reported blogs refer to the user's top five surfing locations.

Figure 4 **Probability That Users i and j Form a Link Against $D(i, j)$ by Different Distances Between Users' Homes (x), Where $D(i, j)$ Is the Distance Between i's Favorite Surf Location and j's Home Location**



equation for identification. We first present an identification strategy using an IV assumption that delivers point identification of the effect of interest in combination with a dynamic panel model analogous to Equation (1). The IV assumption uses the number of friend requests to an agent's friends as an instrument for his friend requests. This strategy uses the variation in the network position of users for identification and is broadly related to a recent literature that emphasizes the role of group-size variation in social networks for identification of endogenous social effects (e.g., Lee 2007, Bramoullé et al. 2009, Oestreicher-Singer and Sundararajan 2012). The IV assumption can, however, be violated if friends receive tie-formation requests in a way that is directly related to the focal agent's blogging activity (e.g., many agents arrive on i's webpage on Soulrider.com because he or she blogs a lot, and those that arrive on i's webpage then send tie requests to i's friends, thus causing those tie requests to be directly a function of i's blogging). Although fixed effects for i address this concern to some extent, it does not accommodate the time variation in the user's friendship formation. Subsequently, we also present estimates from two approaches that utilize alternative monotonicity assumptions and deliver bounds on the treatment effects.

Our IV model uses friendship requests received by the friends of each focal agent as instruments for *friends* in a linear model analogous to Equation (1):

$$b_{it} = \alpha_{2i} + \gamma_{2t} + \theta_2 f_{it} + \lambda_2 \mathscr{B}_{it} + \zeta_2 \mathscr{F}_{it} + \delta_2 z_{it} + \varepsilon_{2it}. \quad (2)$$

To understand the friend-of-friend instruments, suppose there are only three agents: the focal agent i and agents j and l. Further suppose that i and j are connected (i.e., already friends), but l is not connected to either. Let $\mathbb{I}_{l\to i,\,t}$ be an indicator of whether agent l sends a friendship request to connect with the focal agent i in period t. Then the number of friend requests i receives is $f_{it} = \mathbb{I}_{l\to i,\,t}$. From Equation (2), because f_{it} affects b_{it}, $\mathbb{I}_{l\to i,\,t}$ has a direct effect on b_{it}. Now suppose l decides to send a friend request to i's friend j; i.e., $\mathbb{I}_{l\to j,\,t} = 1$. Clearly, $\mathbb{I}_{l\to j,\,t}$ is correlated with the endogenous variable $f_{it} = \mathbb{I}_{l\to i,\,t}$ because factors that influenced l to send a request to i could also influence him to send a request to i's friend j. However, as $\mathbb{I}_{l\to j,\,t}$ is not a request to i, $\mathbb{I}_{l\to j,\,t}$ does not have a direct effect on i's blogging, b_{it}. Hence, $\mathbb{I}_{l\to j,\,t}$ serves as an instrument for f_{it}. Essentially, we use the requests received by i's friends as instruments for the requests he receives.

To operationalize the variable, we define fof_{it} (friends of friends), collecting the set of friends i has in $\mathscr{A}_t(i)$ and then counting the number of friend requests received by i's friends in period t as $fof_{it} = \sum_{j\in\mathscr{A}_t(i),\,k\notin\mathscr{A}_t(j)} \mathbb{I}_{k\to j,\,t}$. We expect *fof* to be positively correlated with *friends* through common characteristics that dissipate as a function of the network distance. For example, immediate friends presumably have similar levels of "gregariousness," which in turn leads to similar rates of friendship formation. Exclusion of *fof* from the *blogs* equation is based on the premise that user i does not set his blogging output in response to friendship requests received by anyone other than himself. Intuitively, we exploit variation in network position under the assumption that after controlling for any common shocks (via agent and time fixed effects), an agent's content generation is likely driven only by the environment facing him, and not by his friends. As with the wind data, observations of *fof* are of daily frequency, whereas the panel frequency is weekly. We operationalize the instrument as the weekly mean of each agent's *fof* observations. That is, we compute $fof\_mean_{it} = mean(fof_{id})$, where d indexes the daily observations corresponding to week t. Summary statistics of *fof_mean* are provided in Table 2.

3.2.1. Identification with Weaker Assumptions. If the IV assumptions above are violated, the strategy outlined above will not deliver consistent estimates of the causal effect of *blogs* and *friends* on each other. We now consider what can be learned about the causal effects under much weaker assumptions that do not assume exclusion restrictions. The trade-off associated with the weaker assumptions will be that we will be able to obtain only bounds and not point estimates of the effects.

We assume that the *blogs* is monotonic in *friends* and that *friends* is monotonic in *blogs*. Following Manski (1997) and Manski and Pepper (2000), we combine this monotone treatment response (MTR) assumption with a monotone treatment selection (MTS) assumption to present sharp bounds on the treatment effects. Subsequently, we utilize an alternative identification strategy using wind and *fof* as MIV in the sense of Manski and Pepper (2000). This latter strategy delivers a separate set of bounds on the treatment effect. We find that our results are broadly consistent across these approaches. We provide a short discussion of the bounding strategies below, pointing the reader to Manski (1997) and Manski and Pepper (2000) for more details on these estimators.

MTR + MTS. We provide a brief discussion of bounding the effect of *friends* on *blogs* (the discussion holds analogously for obtaining the effect of *blogs* on *friends*). Using the terminology of Manski (1997), *blogs* (b) is the outcome of interest while *friends* (f) is the treatment, and the response function for an individual i is $b_i(f)$, which is also allowed to be heterogeneous across users. The goal of estimation is to recover $\mathbb{E}[b_i(f)]$, the mean response of *blogs* to *friends* level f in the population. Knowledge of the mean response is sufficient to determine the average treatment effect (ATE), the causal effect on blogging of increasing *friends* from f_1 to f_2, $\text{ATE} = \Delta(f_1, f_2) = \mathbb{E}[b_i(f_2)] - \mathbb{E}[b_i(f_1)]$.

The data provide the joint distribution of realized friend requests and blogging responses. We denote the *realized* treatment levels (friendship requests) for agent i in the data as $\mathfrak{f}_i$ and the blogs generated by i given $\mathfrak{f}_i$ as $b_i(\mathfrak{f}_i)$; $\{\mathfrak{f}_i, b_i(\mathfrak{f}_i)\}$ is observed in the data. A *conjectured* treatment level is denoted f. The blogging responses to the conjectured treatment level, $b_i(f)$, are latent outcomes that need to be inferred from the observed information in the data. Calculating the ATE from the data is complicated by two issues. First, observing $\{\mathfrak{f}_i, b_i(\mathfrak{f}_i)\}$ is insufficient to calculate $\mathbb{E}[b_i(f)]$ for a conjectured treatment level of friends f, because we do not observe $b_i(f)$ for agent i when $f \neq \mathfrak{f}_i$.[10] Second, the observed treatment levels $\mathfrak{f}_i$ are not randomly allocated across i. We expect those who tend to blog more to attract more friend requests; hence, the observed treatment levels f_i are not statistically independent of response functions, $b_i(\mathfrak{f}_i)$. The conditional distribution of the observed blogging functions at the realized levels of friendship requests are therefore not informative of the response functions at other levels of friendship requests. This selection on unobservables results in a selection bias problem, which prevents us from calculating the ATE without further assumptions. This situation is the root of the empirical identification problem.

We follow Manski (1997) and Manski and Pepper (2000) and use two intuitive restrictions on the blogging response functions (and analogously, friending response functions). Later, we show these assumptions are not rejected by our data:

Definition 1 (Monotone Treatment Response). For each i, $f_2 \geq f_1 \Rightarrow b_i(f_2) \geq b_i(f_1)$.

Definition 2 (Monotone Treatment Selection). For each i, $\mathfrak{f}_2 \geq \mathfrak{f}_1 \Rightarrow \mathbb{E}[b_i(f) \mid f = \mathfrak{f}_2] \geq \mathbb{E}[b_i(f) \mid f = \mathfrak{f}_1]$.

The MTR and MTS assumptions imply weak monotonicity on the response functions and average response respectively. Although both are consistent with the statistical fact that "more popular agents blog more," both interpret this outcome in different ways. The MTR assumption says that all things held equal, blogging increases in the conjectured levels of friendship requests received by each agent. The MTS assumption asserts that agents who have higher numbers of friendship requests have weakly higher mean blogging functions than those who receive fewer friendship requests. Both are consistent with prior expectations about user behavior here. Manski and

[10] Although panel data on individuals are available, the presence of feedback effects and lagged actions in the model determining outcomes implies that the user in the future is not comparable to the user in the present.

Pepper (2000) show that combining MTR and MTS assumptions yields the following bounds on the mean response:

$$\sum_{\mathfrak{u}<f} \mathbb{E}[b_i \mid \mathfrak{f}=u]P(\mathfrak{f}=u)+\mathbb{E}[b_i|\mathfrak{f}_i=f]P(\mathfrak{f}_i\geq f)\leq \mathbb{E}[b_i(f)]$$
$$\leq \sum_{\mathfrak{u}>f} \mathbb{E}[b_i|\mathfrak{f}=u]P(\mathfrak{f}=u)+\mathbb{E}[b_i \mid \mathfrak{f}_i=f]P(\mathfrak{f}_i\leq f). \quad (3)$$

These bounds are sharp in the sense that in the absence of any additional assumptions, every possible value within the upper and lower bounds in Equation (3) has an equal chance of being the mean response in the population. Note that nothing is assumed about the process of treatment selection, and no across-agent restrictions on the response of blogging to friends are imposed. Hence, these bounds are robust to nonrandom selection and to functional form (and, in particular, they allow for nonlinear response). All terms in the bounds in Equation (3) can be estimated nonparametrically from the data.[11]

The bounds in (3) in turn imply bounds on the ATE, $\Delta(f_1,f_2)$. By MTR, $b_i(f_2)\geq b_i(f_1)$ if $f_2\geq f_1$, implying the lower bound on the ATE is 0. The upper bound on the ATE is the upper bound on $\mathbb{E}[b_i(f_2)]$ minus the lower bound on $\mathbb{E}[b_i(f_1)]$:

$$\begin{aligned}\Delta(f_1,f_2)&\\ \leq \sum_{u<f_1}&\big(\mathbb{E}[b_i \mid f=f_2]-\mathbb{E}[b_i \mid f=u]\big)P(f=u)\\ &+\big(\mathbb{E}[b_i \mid f=f_2]-\mathbb{E}[b_i \mid f=f_1]\big)P(f_1\leq f\leq f_2)\\ &+\sum_{u>f_2}\big(\mathbb{E}[b_i \mid f=u]-\mathbb{E}[b_i \mid f=f_1]\big)P(f=u). \quad (4)\end{aligned}$$

MTR + MIV. Finally, we also consider alternative bounds under an MIV assumption. The concern with the IVs is that they could have a direct effect on outcomes and thereby violate the exclusion conditions. For instance, we were worried that wind could directly effect *friends* (and not just indirectly as an IV for blogging) and that *fof* could directly affect blogging (and not just indirectly as an IV for *friends*). We now consider a class of weaker assumptions that allows us to utilize the variation provided by wind and *fof*, while accommodating the fact that the exclusion restrictions do not hold. Following Manski and Pepper (2000), we assume that *fof*, denoted z, is an MIV for *friends* (analogously that wind is an MIV for *blogs*):

Definition 3 (Monotone Instrumental Variable). Covariate z is an MIV if for all $z_2\geq z_1 \Rightarrow \mathbb{E}[b_i(f) \mid z=z_2]\geq \mathbb{E}[b_i(f) \mid z=z_1]$.

We can contrast the MIV assumption with an assumption that z is an IV: that for all (z_2,z_1), $\mathbb{E}[b_i(f) \mid z=z_2]=\mathbb{E}[b_i(f) \mid z=z_1]$. Thus, to impose that *fof* is an IV implies asserting an exclusion restriction that agents connected to those who obtain similar numbers of friendship requests (*fof*) have the same mean blogging propensity. This may be violated, as previously explained. However, to impose that *fof* is an MIV weakens the exclusion restriction to a monotonicity restriction that agents who are connected to those who receive more friendship requests have weakly higher blogging propensity than agents connected to those who receive fewer friendship requests. It is reasonable to expect that all things being equal, agents tend to blog more (and not less) when they are connected to more friends. Similarly, to impose that wind is an MIV weakens the restriction that wind does not directly affect friends to a much weaker monotonicity restriction, that agents who surf at windier locations receive more friendship requests than agents who surf at less windy locations. These MIV assumptions are thus fairly reasonable for Soulrider.com and in line with what we expect a priori.

Following Manski and Pepper (2000), combining MIV and MTR assumptions provides the following sharp bounds on $\mathbb{E}[b_i(f)]$:[12]

$$\sum_{z\in\mathfrak{Z}} P(v=z)\Big\{\sup_{u_1\leq z}\big[\mathbb{E}[b_i \mid v=u_1, f\geq \mathfrak{f}_i]P(f\geq \mathfrak{f}_i \mid v=u_1)$$
$$+\underline{\mathfrak{B}}P(f<\mathfrak{f}_i \mid v=u_1)\big]\Big\}\leq \mathbb{E}[b_i(f)]$$
$$\leq \sum_{z\in\mathfrak{Z}} P(v=z)\Big\{\inf_{u_2\geq z}\big[\mathbb{E}[b_i \mid v=u_2, f\leq \mathfrak{f}_i]P(f\leq \mathfrak{f}_i \mid v=u_2)$$
$$+\bar{\mathfrak{B}}P(f>\mathfrak{f}_i \mid v=u_2)\big]\Big\}. \quad (5)$$

Intuitively, the bounds impose restrictions on the mean response implied by an MTR assumption, conditional on a given value of the MIV, and then find either the maximum or the minimum of these restrictions over all possible values taken by the MIV. As in the previous case, all values of the lower and upper bounds are observed in the data, enabling us to estimate bounds on the conditional mean nonparametrically. Bounds on the treatment effect can then be computed from the above bounds on the conditional mean, as in Equation (4).

This concludes our discussion of identification. In our results section, we report estimates for the models without instruments and for treatment effects under the IV assumption, the MTR + MTS assumption, and the MIV + MTR assumptions.

[11] Bounds using either MTS or MTR alone, which are derived in Manski and Pepper (2000), are wider and less informative.

[12] In formula (5), $\underline{\mathfrak{B}}$ ($\bar{\mathfrak{B}}$) is the minimum (maximum) of the realized value of blogs (b).

4. Data and Results

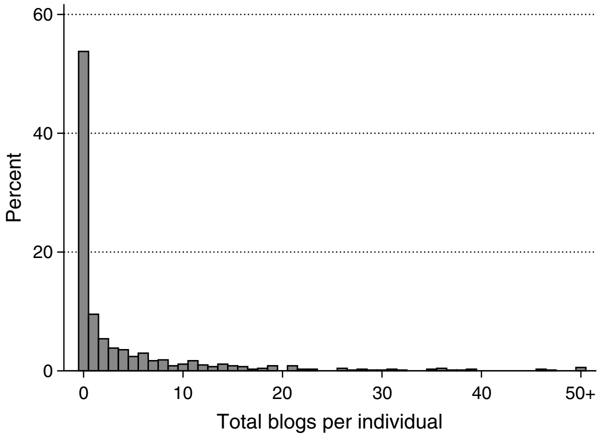
We now discuss the data in more detail, beginning with some stylized patterns. First, we describe key patterns in the generation of content and the pattern of social ties. Then, we check for interrelationships in content generation and network structure.

Table 2 presents the descriptive statistics for the *blogs* and *friends* variables. On average, users post about 0.05 blogs per week (maximum 7), and add about 0.03 friends per week (maximum 6). There are also a large number of user-weeks when no blogs are posted or no friends are added. An average user week starts with about 5 prior blogs posted and about 10 friend requests already received. We also add the variable *days_online* to our specification, which counts the number of days since the user registered on Soulrider.com. We conjecture that this variable may help inform the rate of content posting, because new users may be more likely to be more active than existing users. From Table 2, the average user has been registered on Soulrider.com for about 909 days. Figure 5 presents histograms of the total content posting across users and the total friendship requests across users in the data set. Figure 5 documents significant heterogeneity across users in both their content-generation behavior and in the extent to which they receive requests to form online ties on Soulrider.com. The cross-sectional correlation across individuals in total blogging and total friends requests is positive and significant ($\rho = 0.35$).

Table 3 **Joint Distribution of *friends* and *blogs***

	friends				
blogs	0	1	2	3+	Total
0	53,744	1,039	70	10	54,863
1	1,470	113	26	6	1,615
2	377	43	8	2	430
3	72	9	2	0	83
4+	41	6	4	0	37
Total	55,704	1,208	110	18	57,040

Figure 5 **Histograms of Total Blogs and Friends Across Users**



The joint distribution of *friends* and *blogs* is summarized by the cross-tabulation in Table 3. There is a large mass point at zero. The across-consumer, across-time correlation between the variables is positive and significant ($\rho = 0.11$).

4.1. Results

4.1.1. The Effect of *blogs* on *friends*. Table 4 presents results from the estimation of the linear model

Table 4 **GMM Estimation of Linear Panel Model:** *friends*(*blogs*)

	OLS	IV†	FE	IV†	FE	IV†	FE	IV†
blogs (*b*)	0.056***	0.142***	0.055***	0.154***	0.043***	0.124**	0.042***	0.151**
	(0.009)	(0.033)	(0.009)	(0.042)	(0.007)	(0.041)	(0.007)	(0.055)
prior_friends (*F*)	0.001***	0.001***	0.002***	0.001***	−0.003	−0.003	−0.003	−0.003
	(0.000)	(0.000)	(0.000)	(0.000)	(0.006)	(0.006)	(0.006)	(0.006)
prior_blogs (*B*)	−0.000***	−0.001***	−0.000***	−0.001***	0.002	0.004*	0.003	0.004*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)	(0.001)	(0.002)
days_online	−0.000***	−0.000***	−0.000***	−0.000**	−0.000*	−0.000**	−0.001	0.002
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.003)
fof_mean	0.029***	0.027***	0.028***	0.025***	0.025***	0.025***	0.025***	0.024***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	0.019***	0.016***	0.014***	0.013***				
	(0.002)	(0.002)	(0.003)	(0.003)				
Overid χ^2		0.217		1.188		0.586		0.117
Overid p-value		0.641		0.276		0.444		0.732
Weak ID F-statistic		54.422		40.337		49.539		34.072
Week fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Individual fixed effects	No	No	No	No	Yes	Yes	Yes	Yes

Notes. Dependent variable is friends per week. Ordinary least squares, bi level fixed effects, and fixed effects IV models reported. Dynamic panel specification estimated via GMM. Robust standard errors clustered at the user level reported in parentheses.

†Instruments: *fof_mean, lagged prior_friends, lagged prior_blogs, days_online, wind_mean, wind_sd.*

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

measuring the effect of *blogs* on *friends*, in which we also add past blogging, past friendship requests received, and the number of days on the website as additional variables explaining friendship requests received in the current period. For consistency with our first identification strategy for the reverse equation, we also include the *fof* variable as a covariate in these regressions. Looking at the first column of Table 4, we see there is preliminary evidence of a feedback effect: The effect of *blogs* on *friends* is strongly statistically significant. However, these estimates are likely upward biased because of the lack of control for homophily or correlated unobservables. The table also presents results including week fixed effects to control for time-varying unobservables that drive blogging and friendship formation. We see that week fixed effects do not change the estimates much, suggesting that such common unobservables may not be first order for these data. Nevertheless, the direction of the change in the estimate is consistent with our intuition: Once we control for this spurious source of correlation, we expect the parameter on *blogs* to decrease in magnitude. Referring to Table 4, we see this is indeed the case.

We now discuss the results from adding individual fixed effects into the previous specification. Note that this specification is very demanding of the data, as the inclusion of both individual and week fixed effects implies that all variation common to individuals within a given week, as well as variation common to weeks for a given individual, are fully controlled for and are not used to inform the causal effects. The fixed effects control for homophily and are expected to correct an upward bias in the estimation of the causal effects. Looking at the last column of Table 4, we see that this is indeed the case. Looking at the first and fifth columns, we see that the effect of *blogs* on *friends* has dropped from 0.056 to 0.042 when adding user fixed effects (a 25% decrease). We also see that individual fixed effects control for a large source of unobserved within-user persistence in the data. For testing statistical significance, note that all tables report robust standard errors that have been clustered at the user level.

Table 4 also presents estimates of the model when instrumenting for *blogs* using the wind speed variables. We first discuss whether these instruments are working correctly. Note that, given the simultaneous equations setup, the first stage for the IV in this equation is essentially the reverse regression, the effect of *friends* on *blogs*, which is presented in Table 5. There, we see the wind variables are significant in explaining blogging. Then we discuss whether we are subject to a weak instruments problem. To test for weak instruments, we report the Kleibergen and Paap (2006) *rk*-statistic, which is a generalization of the weak IV test to the case of non-i.i.d. (independent and identically distributed) errors. The null is that the instruments are weak, and a rough thumb rule for empirical work is that there is no weak instruments problem if the *rk*-statistic is >10. Looking at the columns in Table 4, we see that this is the case: the weak identification statistics are all greater than 30. Furthermore, we see that the overidentifying restrictions for the instruments are not rejected in any of the models, and the fit is good. Overall, these diagnostics indicate that the instruments are working properly.

Table 5 GMM Estimation of Linear Panel Model: *blogs*(*friends*)

	OLS	IV†	FE	IV†	FE	IV†	FE	IV†
friends (f)	0.162^{***}	0.776^{***}	0.150^{***}	0.632^{***}	0.087^{***}	0.393^{***}	0.095^{***}	0.347^{**}
	(0.022)	(0.130)	(0.021)	(0.135)	(0.015)	(0.113)	(0.015)	(0.114)
prior_friends (F)	0.000	-0.002^{***}	0.000	-0.001^{**}	-0.020^{*}	−0.015	0.004	0.006
	(0.001)	(0.001)	(0.001)	(0.001)	(0.008)	(0.008)	(0.007)	(0.006)
prior_blogs (B)	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	-0.019^{***}	-0.019^{***}	-0.015^{***}	-0.015^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.005)	(0.004)	(0.004)
days_online	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	0.000^{***}	0.000^{***}	-0.035^{***}	-0.034^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.006)	(0.006)
wind_mean	0.045^{***}	0.040^{***}	0.037^{***}	0.035^{***}	0.048^{***}	0.045^{***}	0.057^{***}	0.054^{***}
	(0.007)	(0.007)	(0.011)	(0.010)	(0.006)	(0.006)	(0.010)	(0.009)
wind_sd	0.034^{***}	0.031^{***}	0.048^{***}	0.041^{***}	0.038^{***}	0.039^{***}	0.040^{***}	0.038^{***}
	(0.010)	(0.008)	(0.011)	(0.010)	(0.009)	(0.008)	(0.010)	(0.010)
Constant	-0.076^{***}	-0.077^{***}	-0.097^{***}	-0.092^{***}				
	(0.014)	(0.013)	(0.019)	(0.019)				
Weak ID F-statistic		102.239		90.073		65.235		61.381
Week fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Individual fixed effects	No	No	No	No	Yes	Yes	Yes	Yes

Notes. Dependent variable is blogs per week. Ordinary least squares, bilevel fixed effects, and fixed effects IV models reported. Dynamic panel specification estimated via GMM. Robust standard errors clustered at the user level reported in parentheses.

†Instruments: *fof_mean, lagged prior_friends, lagged prior_blogs, days_online, wind_mean, wind_sd.*

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

We also see that the *fof* variable is significant and relevant in driving friendship requests, which is important for the identification of the reverse equation. Furthermore, the stock variables, past blogging, past friendship requests, and days on Soulrider.com are significant in the ordinary least squares regressions, but they are not significant once time and individual-specific fixed effects are included. This result persists when we use instruments.

Finally, we see that the magnitude of the coefficient on *blogs* in the *friends* equation has increased after instrumenting. We found this result surprising, as, a priori, we expected the bias to go in the other direction. However, this result is plausible and can be explained by stories where the unobservables are negatively correlated with the endogenous variable. For instance, we conjecture that unobservables that drive friendship formation could proxy for extroversion, friendliness, or windsurfing skill (more users want to be friends with better windsurfers, all things held equal). If extroverts post more blogs, the unobservables would be positively correlated with blogs, and we would observe an upward bias. If, on the other hand, extroverts tend to spend more time offline and post fewer blogs, then unobservables would be negatively correlated with blogging and we would observe a downward bias (as we see here). We do not take a strong stance on what these unobservables represent, because either story seems reasonable.

4.1.2. The Effect of *friends* on *blogs*. Table 5 presents the results from linear panel data models of the effect of *friends* on *blogs*. The dependent variable in the regression is *blogs*. Analogous to the estimation of the previous model, we add past blogging, past friendship requests received, and the number of days on the website as controls. Consistent with the simultaneous equation framework in which we used wind speeds as instruments for *blogs* in the reverse equation, we include the wind speed instruments as covariates in these regressions. We also report estimates from the IV strategy in which we use friends of friends (*fof_mean*) as an IV for friendship requests. Results with and without user and week fixed effects and with and without instrumenting are reported.

From Table 5 we find that the friendship requests have a statistically significant effect on blogging. Mirroring the pattern of results from the regressions in Table 4, we find that controls for individual and week fixed effects tend to reduce the effect of friendship requests, and instrumenting increases the magnitude. We find some evidence of negative state dependence in blogging, with new blogs declining over time for those who have been on the website for a long time and those who have already posted a lot of content. We also find that the wind speed variables are strongly statistically significant in explaining blogging. The propensity to post content is increasing in the mean wind speed at the user's preferred surfing locations (presumably reflecting increased surfing activity) and in the variation in the wind speeds at the preferred surf location (i.e., the fact that the wind is too low to facilitate surfing also has informational value and may result in posts). Table 5 also reports diagnostics on weak instruments for the friends-of-friends instrument, which indicates no concern of a weak instruments problem. As this equation is just identified, we do not report tests of overidentifying restrictions.

4.1.3. Bounds Estimates: MIV, MTR, and MTS Strategies. We now present bounds estimates of the effects under the MIV, MTR, and MTS assumptions discussed previously. Compared to the linear IV previously, these estimates are nonparametric and impose no functional form restrictions on the nature of the relationship between *friends* and *blogs*. The fact that the treatment variables are discrete facilitates nonparametric estimation using a bin estimator for the lower and upper bounds (i.e., we do not need to do any smoothing). In Table 6 we report bounds

Table 6 **Bounds on the Effect of *friends* on *blogs*, as Well as the Average Treatment Effect Using MIV, MTR, and MTS Assumptions; *fof* (Friends of Friends) Used as an MIV**

	Bounds on $\mathbb{E}[blogs \mid friends]$							
	MIV-MTR		MTS-MTR		Upper bound on ATE, $\Delta(f_1, f_2)$			
friends	Lower	Upper	Lower	Upper	f_1	f_2	MIV-MTR	MTS-MTR
0	0.046	0.051	0.047	0.051	0	1	5.747	0.153
	(0.001)	(0.001)	(0.001)	(0.001)			(0.017)	(0.016)
1	0.050	5.793	0.050	0.199	1	2	5.925	0.528
	(0.001)	(0.017)	(0.001)	(0.016)			(0.006)	(0.088)
2	0.052	5.976	0.051	0.578	2	3+	5.944	0.515
	(0.001)	(0.006)	(0.001)	(0.088)			(0.002)	(0.174)
3+	0.052	5.995	0.051	0.566				
	(0.001)	(0.002)	(0.001)	(0.174)				

Note. Standard errors in parentheses, based on 1,000 bootstrap replications.

Table 7 **Bounds on the Effect of *blogs* on *friends*, as Well as the Average Treatment Effect Using MIV, MTR, and MTS Assumptions; *wind_mean* (Wind Speed) Used as an MIV**

	Bounds on $\mathbb{E}[friends \mid blogs]$							
	MIV-MTR		MTS-MTR		Upper bound on ATE, $\Delta(b_1, b_2)$			
blogs	Lower	Upper	Lower	Upper	b_1	b_2	MIV-MTR	MTS-MTR
0	0.022	0.025	0.022	0.026	0	1	4.846	0.091
	(0.001)	(0.001)	(0.001)	(0.001)			(0.430)	(0.009)
1	0.025	4.868	0.026	0.113	1	2	5.927	0.124
	(0.001)	(0.430)	(0.001)	(0.009)			(0.299)	(0.021)
2	0.026	5.952	0.026	0.150	2	3	6.337	0.139
	(0.001)	(0.299)	(0.001)	(0.021)			(0.264)	(0.044)
3	0.026	6.363	0.026	0.165	3	4+	6.709	0.226
	(0.001)	(0.264)	(0.001)	(0.044)			(0.155)	(0.091)
4+	0.026	6.735	0.026	0.252				
	(0.001)	(0.155)	(0.001)	(0.091)				

Note. Standard errors in parentheses, based on 1,000 bootstrap replications.

on the average treatment effect of *friends* on *blogs* combining MIV and MTR assumptions in one strategy and combining MTR and MTS assumptions in another (bounds from only one proved to be wider than those reported here, as expected). In the first four panels of Table 6, we report the mean response of blogging to various levels of the treatment, *friends*. For instance, the first row reports on $\mathbb{E}[b_i(f) \mid f = 0]$. Upper and lower bounds on $\mathbb{E}[b_i(f)]$ for each level of f are presented under the two monotonicity assumptions. Standard errors for the upper and lower bounds, which are computed using a nonparametric panel bootstrapping procedure, are reported in parentheses. The bounds intervals are estimated precisely. Furthermore, we see the mean response is increasing in the treatment levels, consistent with the monotonicity assumption. This is also a test for whether the monotonicity conditions hold; the table indicates the data do not reject monotonicity.

An intuitive way to interpret the intervals is to compute the ATE, reported in the right panel of the table. As mentioned before, the lower bound on the ATE is zero under the MTR assumption. Therefore, the implied upper bounds are presented. We see that the MIV-MTR assumptions imply wide intervals on the ATE: for instance, increasing *friends* from 0 to 1 would on average increase blogging anywhere from 0 to 5.747 posts per week. These bounds are relatively uninformative, because *any* value in this range has an equal chance of being the ATE.[13] However, we see that the MTR-MTS assumptions have substantial bite in bounding the treatment effect. Under these assumptions, the upper bound on the average effect on blogging of increasing *friends* from 0 to 1 is 0.153 per week, which is significantly smaller and much more informative. Given this, our preferred estimator is the one that uses the MTR-MTS assumption. Comparing to Table 5, we see the effects reported for the marginal effect of friends under fixed effects panel IV models (0.347) is in the range of the bounds implied by MTR-MTS (and the wider less-informative bounds implied by MIV-MTR).

In Table 7 we report bounds on the average treatment effect of *blogs* on *friends*, combining MIV and MTR assumptions in one strategy as before and combining MTR and MTS assumptions in another. We use wind as an MIV for blogging. In the first four panels of Table 7, we report the mean response of friending to various levels of the treatment, *blogs*, as well as the average treatment effect from changing blogs. We see that the MIV-MTR assumptions imply wide intervals on the ATE, as in the previous case, but that the MTR-MTS assumptions are effective in informatively bounding the treatment effect. Thus, our preferred estimator for this side of the equation is also the one that uses the MTR-MTS assumption. As before, we see that monotonicity is not rejected. Again, comparing to Table 4, we see the effects reported for the marginal effect of blogs under fixed effects panel IV models (0.151) is in the range of the bounds implied by MTR-MTS (as well the wider, less-informative bounds implied by MIV-MTR).

These results, obtained across several alternative assumptions and identification strategies, suggest that our estimates are reasonably robust. To interpret our estimates, we can roughly convert into revenue terms using some back-of-the-envelope arithmetic. As with most social networking sites, Soulrider.com's primary revenue stream is from online advertising, and thus

[13] The wide bounds in our application are a result of the high degree of skew in the conditional distribution of the treatment (f) at all levels of the instrument (z). The large point-mass at the zero friends treatment level that implies that $P(f > \mathfrak{f}_i \mid v) \approx 1$ for $\mathfrak{f}_i > 0$ at all levels of the MIV, which in turn implies that the upper bound in Equation (5) will approximately equal the maximal response value.

incremental revenue from any intervention is realized by generating additional page views, which in turn depend on the amount of user-generated content on the site. To estimate a predictive relationship between blog production and page impressions, we can regress the *blogs* variable on the number of page impressions on the user's webpage (t-statistics in parentheses):

$$PI_{it} = a_i + Xb_{it} = a_i + \underset{(54.24)}{368.04}\, b_{it}.$$

Implicitly, we assume in this regression that aspects of the user's page "quality" that affect page-views are controlled for via the fixed effect a_i. Because Soulrider.com receives approximately 0.002 CHF per page impression, we obtain that revenue is related blogging as $R(b) = 0.002 \times 368.0 \times b = 0.736 \times b$ (note, 1 CHF $\approx$ 1 USD). Thus, an incremental blog is worth about 0.736 CHF on average to Soulrider.com. If one takes feedback effects into account, however, we should also incorporate the fact that, on the margin, incremental blogging results in incremental friending, which in turns leads to incremental blogging, and so on, till feedback settles. Using the linear IV estimates, we can compute (by solving (1) and (2) for the reduced form for b) that such local feedback effects contribute about an additional 5.5% revenue on average across users. There is also significant variation across users.[14]

5. Conclusions

This paper adds to an emerging literature relevant to social networking website content management strategies. We empirically document that ties can help facilitate content generation on a site, thereby creating a link between social tie formation and advertising revenues, and generating local network effects between content and ties. We use detailed data from an online social network to conduct our empirical analysis. The data enable us to control for several confounds that have typically plagued empirical analysis of online social interactions. We discuss what assumptions are required in empirical strategies that seek to identify causal effects in this setting. We outline two approaches, one based on exclusion restrictions and the other on monotonicity assumptions. If exclusions or monotonicity are violated, our estimates should be interpreted as correlational and not causal. We present some evidence in support of the exclusion assumptions, but some aspects remain untestable. We show that monotonicity is not rejected. We find evidence consistent with network effects and describe the implications for website revenue.

Our analysis is related to the recent interest in marketing to gain understanding of the role of social interactions using data from online social networks. An empirical challenge in this endeavor is that the network structure is endogenous to the actions taken by agents. Augmenting the model of an agent's actions on the network with a model for the network structure requires solving a formidable network formation game. An alternative approach to this problem adopted here is to conduct inference with an incomplete model of network formation under weak assumptions that deliver informative bounds on the causal effects of interest. This approach seems promising for empirical analysis of a wide variety of social and economic networks.

Several extensions of the current work are possible. Data on page views can help map out a richer picture of the strength of ties between users. Data on ad exposures can help sharpen the link between content and advertising revenues. Also interesting would be further research on understanding the mechanisms and moderators of social influence on the site (e.g., Zhang (2010) notes the differing implications for marketing of observational learning versus information sharing, and Du and Kamakura (2011) note that social influence is moderated by spatial and temporal contiguity). An important contribution would be to develop a better understanding of the utility of users from content creation and tie formation and to endogenize the network structure within a tractable empirical model. We leave these extensions to future work.

Acknowledgments

The authors contributed equally and are listed in reverse-alphabetical order. The authors thank J. P. Ferguson, Chris Forman, Wagner Kamakura, Carl Mela, Song Yao, the review team, and seminar participants at the Fuqua School of Business, Georgia Tech, Stanford University, University of Rochester, and at the European Marketing Academy (2011), Marketing Dynamics (2011), Marketing Science (2010, 2011), Net Institute (2010) and the Quantitative Marketing and Economics (2011) conferences for useful comments. The authors thank the NET Institute (http://www.NETinst.org) for supporting this research.

References

Ahn D-Y, Duan JA, Mela C (2011) An dynamic equilibrium model of user generated content. Working paper, Duke University, Durham, NC.

Albuquerque P, Pavlidis P, Chatow U, Chen K-Y, Jamal Z (2012) Evaluating promotional activities in an online two-sided market of user-generated content. *Marketing Sci.* 31(3):406–432.

Ansari A, Koenigsberg O, Stahl F (2011) Modeling multiple relationships in social networks. *J. Marketing Res.* 48(4):713–728.

Bramoullé Y, Djebbari H, Fortin B (2009) Identification of peer effects through social networks. *J. Econometrics* 150(1):41–55.

Braun M, Bonfrer A (2011) Scalable inference of customer similarities from interactions data using dirichlet processes. *Marketing Sci.* 30(3):513–531.

[14] A prior working version of this paper reported these results, which are available from the authors upon request.

Brothers L, Hollan J, Nielsen J, Stornetta S, Abney S, Furnas G, Littman M (1992) Supporting informal communication via ephemeral interest groups. *Proc. 1992 ACM Conf. Comput.-Supported Cooperative Work* (ACM, New York), 84–90.

Bughin JR (2007) How companies can make the most of user-generated content. *McKinsey Quart.* (3):1–4.

Chamberlain G (1992) Comment: Sequential moment restrictions in panel data. *J. Bus. Econom. Statist.* 10(1):20–26.

Chevalier J, Mayzlin D (2006) The effect of word of mouth online: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Chintagunta P, Venkataraman S, Gopinath S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Ciliberto F, Tamer E (2009) Market structure and multiple equilibria in airline markets. *Econometrica* 77(6):1791–1828.

Corbin JM, Strauss AL (2008) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (Sage Publications, Thousand Oaks, CA).

Dellarocas C (2006) Strategic manipulation of Internet opinion forums: Implications for consumers and firms. *Management Sci.* 52(10):1577–1593.

Dhar V, Chang E (2009) Does chatter matter? The impact of user-generated content on music sales. *J. Interactive Marketing* 23(4):300–307.

Du R, Kamakura W (2011) Measuring contagion in the diffusion of consumer packaged goods. *J. Marketing Res.* 48(1):28–47.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Friese S (2011) Using ATLAS.ti for analyzing the financial crisis data. *Forum: Qualitative Soc. Res.* 12(1):Article 39.

Ghose A, Han SP (2011) An empirical analysis of user content generation and usage behavior on the mobile Internet. *Management Sci.* 57(9):1671–1691.

Godes D, Mayzlin D, Chen Y, Das S, Dellarocas C, Pfeiffer B, Libai B, Sen S, Shi M, Verlegh P (2005) The firm's management of social interactions. *Marketing Lett.* 16(3):415–428.

Goldenberg J, Han S, Lehmann D, Hong JW (2009) The role of hubs in the adoption processes. *J. Marketing* 73(2):1–13.

Haile PA, Tamer E (2003) Inference with an incomplete model of english auctions. *J. Political Econom.* 111(1):1–51.

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–1054.

Hartmann WR, Manchanda P, Nair H, Bothner M, Dodds P, Godes D, Hosanagar K, Tucker C (2008) Modeling social interactions: Identification, empirical methods and policy implications. *Marketing Lett.* 19(3):287–304.

Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *J. Interactive Marketing* 18(1):38–52.

Jackson MO (2008) *Social and Economic Networks* (Princeton University Press, Princeton, NJ).

Katona Z, Sarvary M (2009) Network formation and the structure of the commercial World Wide Web. *Marketing Sci.* 27(5):764–778.

Katona Z, Zubcsek P, Sarvary M (2011) Network effects and personal influences: Diffusion of an online social network. *J. Marketing Res.* 48(3):425–443.

Kleibergen F, Paap R (2006) Generalized reduced rank tests using the singular value decomposition. *J. Econometrics* 133(1): 97–126.

Krippendorff K (2004) *Content Analysis: An Introduction to Its Methodology* (Sage Publications, Thousand Oaks, CA).

Kumar V (2011) Why do consumers contribute to connected goods? A dynamic game of competition and cooperation in social networks. Working paper, Harvard Business School, Boston.

Lee L (2007) Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *J. Econometrics* 140(2):333–374.

Lento T, Welser HT, Gu L, Smith M (2006) The ties that blog: Examining the relationship between social ties and continued participation in the Wallop weblogging system. *WWW.Third Annual Workshop on the Weblogging Ecosystem, Edinburgh, Scotland*.

Manski CF (1997) Monotone treatment response. *Econometrica* 65(6):1311–1334.

Manski CF (2000) Economic analysis of social interactions. *J. Econom. Perspect.* 14(3):115–136.

Manski CF, Pepper JV (2000) Monotone instrumental variables: With an application to the returns to schooling. *Econometrica* 68(4):997–1010.

Mayzlin D, Yoganarasimhan H (2012) Link to success: How blogs build an audience by promoting rivals. *Management Sci.* 58(9):1651–1668.

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Rev. Sociol.* 27(27):415–444.

Moffitt RA (2001) Policy interventions, low-level equilibria, and social interactions. Durlauf SN, Young HP, eds. *Social Dynamics* (MIT Press, Cambridge, MA), 45–82.

Narayan V, Yang S (2007) Modeling the formation of dyadic relationships between consumers in online communities. Working paper, Cornell University, Ithaca, New York. http://ssrn.com/abstract=1027982.

Narayanan S, Nair H (2013) Estimating causal installed-base effects: A bias-correction approach. *J. Marketing Res.* Forthcoming.

Nardi BA, Schiano DJ, Gumbrecht M, Swartz L (2004) Why we blog. *Commun. ACM* 47(12):41–46.

Nov O (2007) What motivates Wikipedians? *Commun. ACM* 50(11):60–64.

Ochoa X, Duval E (2008) Quantitative analysis of user-generated content on the Web. *Proc. First Internat. Workshop on Understanding Web Evolution (WebEvolve2008), Beijing*, 19–26.

Oestreicher-Singer G, Sundararajan A (2012) The visible hand? Effects of recommendation networks in electronic markets. *Management Sci.* 58(11):1963–1981.

Stephen AT, Toubia O (2010) Deriving value from social commerce networks. *J. Marketing Res.* 47(2):215–228.

Trusov M, Bucklin RE, Pauwels K (2009) Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site. *J. Marketing* 73(5):90–102.

Wooldridge J (2002) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

Yoganarasimhan H (2012) Impact of social network structure on content propagation: A study using YouTube data. *Quant. Marketing Econom.* 10(1):111–150.

Zhang J (2010) The sound of silence: Observational learning in the U.S. kidney market. *Marketing Sci.* 29(2):315–335.

Zhang K, Sarvary M (2011) Social media competition: Differentiation with user-generated content. Working paper, INSEAD, Fontainebleau, France.

Zhang X, Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *Amer. Econom. Rev.* 101(4):1601–1615.